Amendments to the Specification

Please amend the last paragraph on page 2 and ending on page 3 as follows:

An alternative method for characterization of transcriptional profiling using the acquisition of short sequence tags from the 3' end of mRNA transcripts has been described by others: "Serial Analysis of Gene Expression", Velculescu et al, Science 1995 270:484; "Using the transcriptome to annotate the genome", Saha et al., ~~Nature Biotech 2002 19: 508~~Nature Biotech 2002 20:508-512; "Generation and analysis of melanoma SAGE libraries: SAGE advice on the melanoma transcriptome", Weeraratna et al, ~~Oncogene 2004 1:11~~Oncogene 2004 23(12):2264-2274. The SAGE method consists in digesting total double stranded cDNA with a 4-bp restriction enzyme that cuts at random positions within the cDNA and ligation of linkers to the restriction fragments located at the most 3' end of the transcript, closest to the polyA sequence. These linkers contain a recognition sequence for a Type IIS restriction enzyme that will cut outside of its recognition sequence to generate a restriction fragment consisting in the linker sequence fused to 10-20 bp sequence of the 3' end of the cellular mRNA. These tags are ligated together, and ditags are amplified by PCR, cloned and sequenced. Determination of the frequency of each tag is used to estimate the relative levels of gene expression for each transcript. The advantage of this method is that it allows for the simultaneous quantitative analysis of large number of transcripts without previous tagging. The disadvantage of this method is that the short sequence tags that are generated (12-14 bp in most cases) do not allow a precise assignment of the tag to a particular genomic locus, which precludes the identification of the gene that is being quantified. Recent improvements in this technology have pushed the length of the tag to 17 bp, followed by 4 bp of constant sequence corresponding to the recognition sequence of the first restriction enzyme which takes the total length of the tag to 21 bp. Analyses of human genome sequence have shown that 75% of 21 bp tags happen only once in the genome and can therefore be uniquely assigned to a single genomic locus. That means that even with the latest improvements, 25% of those tags still cannot be uniquely assigned to a single genomic locus. Moreover, as information is obtained only from the 3' end of each transcript, it does not allow characterizing the alternative spliced forms of transcripts. This is important because alternative splicing is greatly responsible for gene expression complexity and protein diversity. In fact, some genetic diseases and cancers have been related to abnormal alternative splicing. Given the

dearth of available information regarding exon-exon and exon-intron boundaries and the importance of such information, it would be desirable to obtain methods for elucidating a transcriptional profile of a given cell, wherein the methods simultaneously provide sequence information corresponding to the boundaries of exons of the genes encoding the proteins contained within the cell.

Please amend the first paragraph beginning on page 18 and ending on page 19 as follows:

Unless otherwise stated, sequence identity/similarity values provided herein refer to the value obtained using the BLAST 2.0 suite of programs using default parameters. Altschul et a., Nucleic Acids Res. 25:3389-3402 (1997). Software for performing BLAST analyses is publicly available, e.g., through the National Center for Biotechnology-Information ~~(http://www.ncbi.nlm.nih.gov/)~~ (http:// world wide web at ncbi.nlm.nih.gov/). This algorithm involves first identifying high scoring sequence pairs (HSPs) by identifying short words of length W in the query sequence, which either match or satisfy some positive-valued threshold score T when aligned with a word of the same length in a database sequence. T is referred to as the neighborhood word score threshold (Altschul et al., supra). These initial neighborhood word hits act as seeds for initiating searches to find longer HSPs containing them. The word hits are then extended in both directions along each sequence for as far as the cumulative alignment score can be increased. Cumulative scores are calculated using, for nucleotide sequences, the parameters M (reward score for a pair of matching residues; always>0) and N (penalty score for mismatching residues; always<0). For amino acid sequences, a scoring matrix is used to calculate the cumulative score. Extension of the word hits in each direction are halted when: the cumulative alignment score falls off by the quantity X from its maximum achieved value; the cumulative score goes to zero or below, due to the accumulation of one or more negative-scoring residue alignments; or the end of either sequence is reached. The BLAST algorithm parameters W, T, and X determine the sensitivity and speed of the alignment. The BLASTN program (for nucleotide sequences) uses as defaults a word length (W) of 11, an expectation (E) of 10, a cutoff of 100, M=5, N=-4, and a comparison of both strands. For amino acid sequences, the BLASTP program uses as defaults a word length (W) of 3, an expectation (E) of 10, and the BLOSUM62 scoring matrix (see Henikoff & Henikoff (1989) Proc. Natl. Acad. Sci. USA 89:10915).

Please amend the last paragraph on page 30 and ending on page 31 as follows:

Comparison of each sequence tag to a nucleotide sequence database can be performed by any of several means known to operators skilled in the art, such as BLAST analysis. Methods of alignment of sequences for comparison are well-known in the art. Optimal alignment of sequences for comparison may be conducted by the local homology algorithm of Smith and Waterman, *Adv. Appl. Math.* 2:482 (1981); by the homology alignment algorithm of Needleman and Wunsch, *J. Mol. Biol.* 48:443 (1970); by the search for similarity method of Pearson and Lipman, *Proc. Natl. Acad. Sci.* 85:2444 (1988); by computerized implementations of these algorithms, including, but not limited to: CLUSTAL in the PC/Gene program by Intelligenetics, Mountain View, California; GAP, BESTFIT, BLAST, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group (GCG), 575 Science Dr., Madison, Wisconsin, USA; the CLUSTAL program is well described by Higgins and Sharp, *Gene* 73:237-44 (1988); Higgins and Sharp, *CABIOS* 5:151-3 (1989); Corpet et al., *Nucleic Acids Res.* 16:10881-90 (1988); Huang *et al., Computer Appl. Biosci.* 8:155-65 (1992), and Pearson *et al., Methods Mol. Biol.* 24:307-31 (1994). The BLAST family of programs which can be used for database similarity searches includes: BLASTN for nucleotide query sequences against nucleotide database sequences; BLASTX for nucleotide query sequences against protein database sequences; BLASTP for protein query sequences against protein database sequences; TBLASTN for protein query sequences against nucleotide database sequences; and TBLASTX for nucleotide query sequences against nucleotide database sequences. *See* CURRENT PROTOCOLS IN MOLECULAR BIOLOGY, Chapter 19, Ausubel *et al.*, eds., Greene Publishing and Wiley-Interscience, New York (1995). Software for performing BLAST analyses is publicly available, e.g., through the National Center for Biotechnology-Information ~~(http://www.ncbi.nlm.nih.gov/)~~(http:// world wide web at ncbi.nlm.nih.gov/).

Please amend the last paragraph on page 37 and ending on page 38 as follows:

In another optional embodiment, the polynucleotide construct comprising the marker exon includes a polynucleotide encoding a negative or positive selection protein for enrichment of the population prior to sorting. Use of the negative or positive selection will remove from the

population all cells with no integration of the polynucleotide, for example via antibiotic resistance. This provides for enriched populations of target cells to overcome any relative inefficiency of the gene trapping of genomic control elements. Enrichment of gene trapped cells will include the use of drug selection (ex. neo^r, $puro^r$, $hygro^r$, zeo^r , HAT^r etc.), affinity separations to include but not limited to {Ab/Ag or Ab/hapten, biotin/streptavidin, glutathione S-transferase (GST) fusion proteins, Polyhistamine fusion proteins (Invitrogen) , calmodulin-binding peptide tag (Stratagene), c-*myc* epitope tag (peptide seq. EQKLISEEDL) (Stratagene) (SEQ ID NO:10), FLAG epitope tag (peptide seq. DYKDDDDK) (Stratagene) (SEQ ID NO:11), V5 epitope (Stratagene), the Linx™ technology {phenyldiboronic acid [PDBA] and salicylhydroxamic acid [SHA]} (Invitrogen) , adhesion, blocking of adhesion, chemotaxis, block of chemotaxis, etc.}, and/or enrichment by FACS using fluorescent Ab, fluorescent Ag, fluorescent substrates or non-fluorescent substrates that become fluorescent after enzymatic cleavage/activation (A complete listing of common fluorescent probes used for applications disclosed herein can be found in PRACTICAL FLOW CYTOMETRY, 3rd ed., Shapiro, Wiley-Liss (1994); HANDBOOK OF FLOW CYTOMETRY METHODS, Robinson, Wiley-Liss (1993); FLOW CYTOMETRY: A PRACTICAL APPROACH, 2nd ed., Ormerod, IRL Press (1994); CURRENT PROTOCOLS IN CYTOMETRY, Robinson, John Wiley & Sons (2000).